Clustering, rethought

In the previous unit, we introduced methods for clustering analysis. **K-means** puts each document into one of k different clusters. This is clean but lacks nuiance; a document can only be in a single cluster.

Hierarchical clustering allows for a bit more variety. We can say that at one level docs A + B and docs C + D make two different clusters, but farther up the tree these combine to create a larger 4 document cluster.

When documents get longer, as in Project 4, these can still both be too restrictive.

Example

Consider the three following Wikipedia pages. Which two pages would you combine to create the first cluster in a hierarchical clustering model?

Rosa Parks

Jackie Robinson

Babe Ruth

Example

Consider the three following Wikipedia pages. Which two pages would you combine to create the first cluster in a hierarchical clustering model?

The trick is that there is really no right answer here.

Rosa Parks

Jackie Robinson

Babe Ruth

Themes

Here is an alternative!

Rosa Parks

Jackie Robinson

Babe Ruth

Baseball **Civil Rights**

Themes

Here is an alternative!

		Baseball	Civil Rights
Rosa Parks		**0%**	**100%**
Jackie Robinson		**50%**	**50%**
Babe Ruth		**100%**	**0%**

Themes

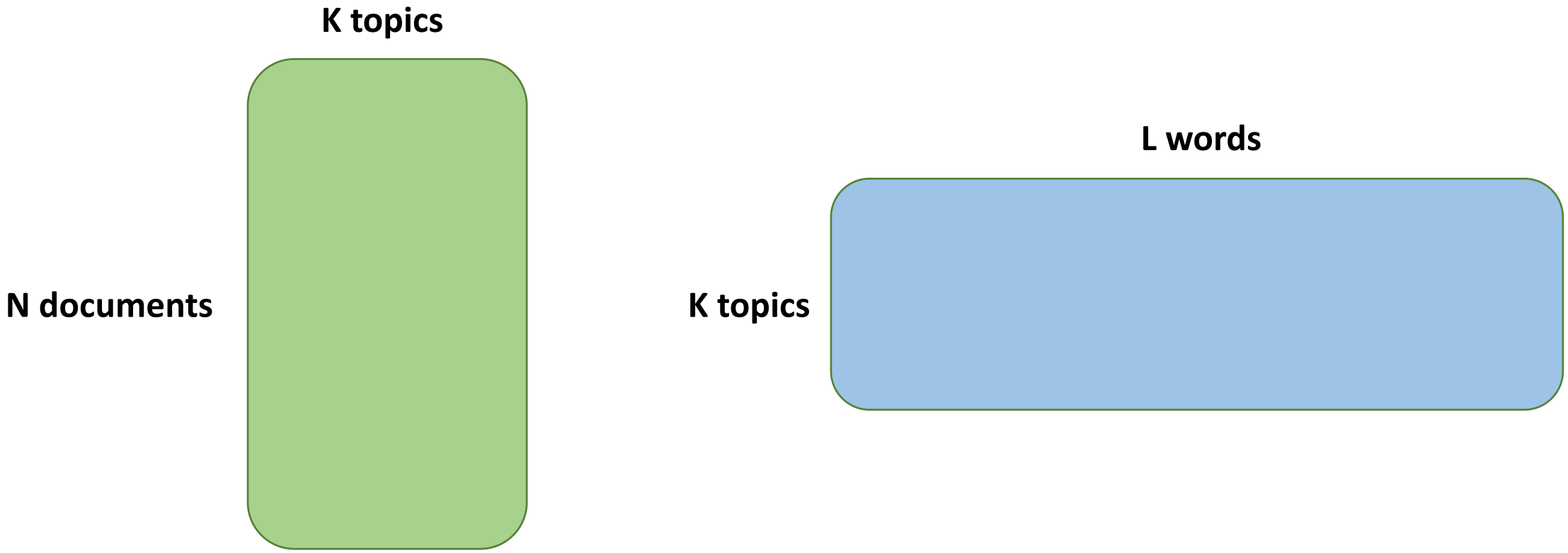
But how to describe the themes?

More probabilities!

	bat	activist	boycott	segregation	ball	homerun	
Baseball	3%	0%	0%	0%	2%	1%	…
Civil Rights	0%	1%	1%	2%	0%	0%	…

Topic Models

Putting this all together, we can describe a **topic model** by two matrices (grids of numbers) describing how documents are distributed over topics and topics are distributed over words.

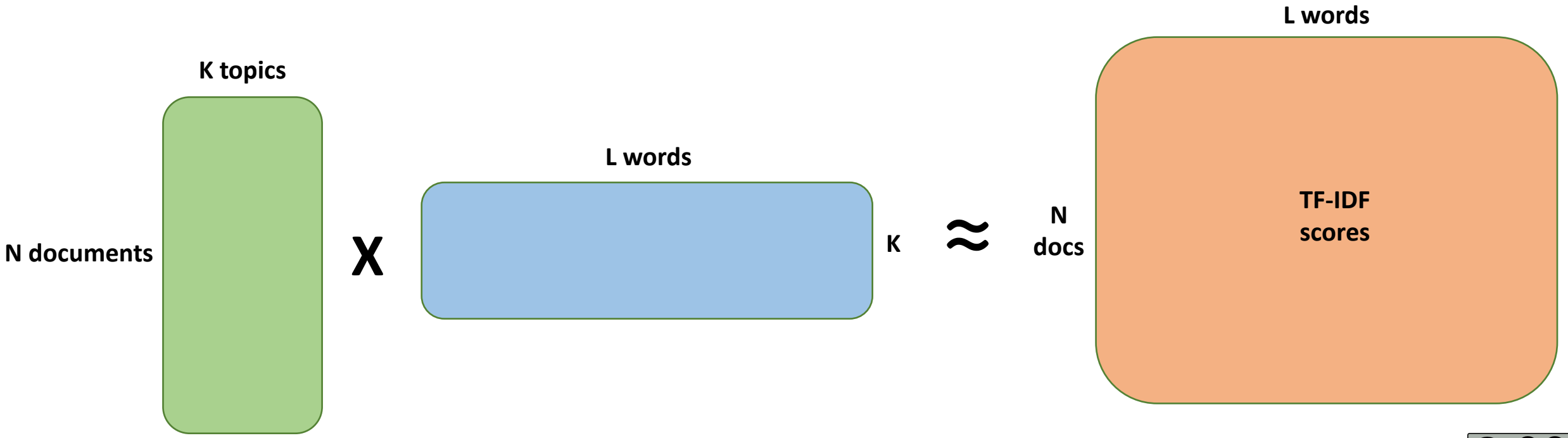


Aside: LSI

One way to find a topic model is to consider the (matrix) product of the two matricies and to try to find values that best approximate the TF-IDF matrix. This is called LSI, or **latent semantic indexing**.

Up to some scaling factors, the green and blue matrices will actually be the first K principal components of the TF-IDF matrix (the blue is the invert = TRUE version for the words).

The technique has some nice applications, but does not produce interpretable topics.



Latent Dirchlet Allocation

The actual technique we will use is a Bayesian method called LDA (latent Dirchlet allocation). The idea is that we find the values for the matrices that maximize the probability of observing the actual data.

Let's see how to do this in R!